Impacts of machine translation and speech synthesis on speech-to-speech translation

Impact of machine translation and speech synthesis on speech-to-speech translation

Kei Hashimoto[a], Junichi Yamagishi[b], William Byrne[c], Simon King[b], Keiichi Tokuda[a]

[a]*Nagoya Institute of Technology, Department of Computer Science and Engineering, Nagoya, Japan.*
[b]*University of Edinburgh, Centre for Speech Technology Research, Edinburgh, United Kingdom*
[c]*Cambridge University, Engineering Department, Cambridge, United Kingdom*

Abstract

This paper provides an analysis of the impacts of machine translation and speech synthesis on speech-to-speech translation systems. The speech-to-speech translation system consists of three components: speech recognition, machine translation and speech synthesis. Many techniques for integration of speech recognition and machine translation have been proposed. However, these techniques have not yet been considered for speech synthesis as well. Therefore, in this paper, we focus on machine translation and speech synthesis, and report a subjective evaluation to analyze the impact of each component. The results of these analyses show that the naturalness and intelligibility of the synthesized speech are strongly affected by the fluency of the translated sentences. In addition, we found that some features correlate well with the average fluency of the translated sentences and the average naturalness of the synthesized speech.

Keywords: speech-to-speech translation, machine translation, speech synthesis, subjective evaluation

1. Introduction

In speech-to-speech translation (S2ST), the source language speech is translated into target language speech. A S2ST system can help to overcome the language barrier, and is essential for providing more natural interaction. A S2ST system consists of three components: speech recognition, machine translation and speech synthesis. In the simplest S2ST system, only the single-best output of one component is used as input to the next component. Therefore, errors of the previous component strongly affect the performance of the next component. Due to errors in speech recognition, the machine translation component cannot achieve the same level of translation performance as achieved for correct text input. To overcome this problem, many techniques for integration of speech recognition and machine translation have been proposed, such as Vidal (1997); Ney (1999). In these, the impact of speech recognition errors on machine translation is alleviated by using N-best list or word lattice output from the speech recognition component as input to the machine translation component. Consequently, these approaches can improve the performance of S2ST significantly. However, these approaches have not yet been considered for speech synthesis as well. The output speech for translated sentences is generated by the speech synthesis component. If the quality of synthesized speech is bad, users will not understand what the system said; the quality of synthesized speech is obviously important for S2ST and any integration method intended to improve the end-to-end performance of the system should take account of the speech synthesis component.

VERBMOBIL is a S2ST project, in the domain of appointment scheduling dialogues, i.e., two persons try to fix a meeting date, time, and place (Noth et al., 2000). In the VERBMOBIL system, the prosodic information extracted from input speech are used for syntactic analysis, semantic construction, dialogue processing, transfer, and speech synthesis. For a better user acceptance, the synthesized output of a translation system should be adapted to the voice of the original speaker. The speech synthesis component of the VERBMOBIL system is only switched to a male or a female voice according to the prosodic information of the original user's utterance. The EMIME project [1] is developing personalized S2ST, such that a user's speech input in one language is used to produce speech output in another language. Speech characteristics of the output speech are adapted to the input speech characteristics using cross-lingual speaker adaptation techniques (Wu et al., 2009). While personalization is an important area of research, this paper focuses on the impact of the machine translation and speech synthesis components on end-to-end performance of an S2ST system. In order to investigate integration methods, we should understand the degree to which each component affects performance. We first conducted a subjective evaluation divided into three sections: speech synthesis, machine translation, and speech-to-speech translation. In this evaluation, various translated sentences were evaluated by using N-best translated sentences output from the machine translation component. The individual impacts of the machine translation and speech synthesis components are analyzed from the results of this subjective evaluation.

The rest of this paper is organized as follows. We will begin

Email addresses: bonanza@sp.nitech.ac.jp (Kei Hashimoto), jyamagis@inf.ed.ac.uk (Junichi Yamagishi), bill.byrne@eng.cam.ac.uk (William Byrne), Simon.King@ed.ac.uk (Simon King), tokuda@nitech.ac.jp (Keiichi Tokuda)

[1] The EMIME project http://www.emime.org/

in Section 2 with a review of related work on integrating natural language generation and speech synthesis for spoken dialog system and integrating machine translation and speech synthesis for S2ST system. In Section 3, the setup of our subjective evaluation is described. The results of analyses between machine translation and speech synthesis are reported in Section 4. Section 5 discusses the objective measures for predict subjective scores. Finally, in Section 6, we conclude with a summary and a discussion of future work.

2. Related work

In the field of spoken dialog systems, the quality of synthesized speech is one of the most important features because users cannot understand what the system said if the quality of synthesized speech is bad. Therefore, integration methods of natural language generation and speech synthesis have been proposed by Bulyko (2002); Nakatsu and White (2006); Boidin et al. (2009).

Bulyko (2002) proposed a integration method of natural language generation and unit selection based speech synthesis which allows the choice of wording and prosody to be jointly determined by the language generation and speech synthesis components. A template-based language generation component passes a word network expressing the same content to the speech synthesis component, rather than a single word string. To perform the unit selection search on this word network input efficiently, weighted finite-state transducers (WFSTs) are employed. The weights of the WFST are determined by join costs, prosodic prediction costs, and so on. In an experiment, this system achieved higher quality speech output. However, this method cannot be used with most existing speech synthesis systems because they do not accept word networks as input.

An alternative to the word network approach is to re-rank sentences from the *N*-best output of the natural language generation component (Nakatsu and White, 2006). *N*-best output can be used in conjunction with any speech synthesis system although the natural language generation component must be able to construct *N*-best sentences. In this method, a re-ranking model selects the sentences that are predicted to sound most natural when synthesized with the unit selection based speech synthesis component. The re-ranking model is trained from the subjective scores of the synthesized speech quality assigned in a preliminary evaluation and features from the natural language generation and speech synthesis components such as word *N*-gram model scores, join cost, and prosodic prediction costs. Experimental results demonstrated higher quality speech output. Similarly, a re-ranking model for *N*-best output was also been proposed by Boidin et al. (2009). In contrast to Nakatsu and White (2006), this model used a much smaller data set for training and a larger set of features, but reached the same performance as reported by Nakatsu and White (2006).

These are integration methods of natural language generation and speech synthesis for spoken dialog systems. In contrast to these methods, our focus is on the integration of machine translation and speech synthesis for S2ST systems. Machine translation output involve some errors: untranslated words, word reordering errors, and wrong lexical choices. However, standard speech synthesis systems are not designed to deal with machine translation errors. In order to handle the errors, Parlikar et al. (2010) proposed some synthesis strategies for unit selection based speech synthesis system: pause insertion, replacing untranslated words with fillers, and using alternative translations from an *N*-best list to tackle bad phonetic joins. In experiments, these synthesis strategies have a positive impact on intelligibility. However, these evaluations were conducted with small data set and small subjects. The detailed analysis with a focus on machine translation and speech synthesis in S2ST was needed in order to investigate integration methods. To this end, we first conducted a large-scale subjective evaluation – using Amazon Mechanical Turk [2] – then analyzed the impact of machine translation and speech synthesis on S2ST.

3. Subjective evaluation setup

3.1. Systems

In the subjective evaluation, a Finnish-to-English S2ST system was used. To focus on the impacts of machine translation and speech synthesis, the correct sentences were used as the input of the machine translation component instead of the speech recognition results. We employed statistical machine translation system and statistical parametric speech synthesis system. In particular to speech synthesis, Wolters et al. (2010) showed that the statistical parametric speech synthesis system was significantly more intelligible than the unit-selection based speech synthesis system.

The system developed in Gispert et al. (2009) was used as the machine translation component of our S2ST system. This system is *HiFST*: a hierarchical phrase-based system implemented with weighted finite-state transducers (Iglesias et al., 2009). For constructing this system, 865,732 parallel sentences from the EuroParl corpus (Koehn, 2005) were used as training data, and 3,000 parallel sentences from the same corpus was used as development data. When the system was evaluated on 3,000 sentences in Gispert et al. (2009), it obtained 28.9 on the BLEU-4 measure.

As the speech synthesis component, HMM-based speech synthesis (Yoshimura et al., 1999; Tokuda et al., 2000) was employed and HTS [3] was used for constructing the speech synthesis system. For constructing this system, 8,129 sentences uttered by one male speaker *Nick*, which was provided by University of Edinburgh and was also used in Wolters et al. (2010), were used for training acoustic models. Speech signals were sampled at a rate of 16 kHz and windowed by an F_0-adaptive Gaussian window with a 5 ms shift. Feature vectors comprised 138-dimensions: 39-dimension STRAIGHT (Kawahara et al., 1999) mel-cepstral coefficients (plus the zero-th coefficient), log F_0, 5 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. We used 5-state left-to-right context-dependent multi-stream MSD-HSMMs (Tokuda

[2] Amazon Mechanical Turk https://www.mturk.com/

[3] HMM-based speech synthesis system (HTS) http://hts.sp.nitech.ac.jp/

et al., 1999; Zen et al., 2004) as acoustic models. Each state had a single Gaussian. Festival [4] was used for deriving full-context labels from the text; the full-context labels include phoneme, part of speech (POS), intonational phrase boundaries, pitch accent, and boundary tones.

3.2. Evaluation procedure

Subjective evaluation was conducted using Amazon Mechanical Turk (AMT). Microtask platforms such as AMT are increasingly used to create speech and language resources (Callison-Burch and Dredze, 2010). AMT provide a welcome link between experimenter and participant. People who are registered with AMT are paid small amounts of money to perform short and simple tasks. Although crowd-workers are cheaper, they are not always reliable (Snow et al., 2008). Recently, investigation whether AMT can be used for comparing the intelligibility of speech synthesis systems has been reported (Wolters et al., 2010). They conducted experiments for comparing intelligibility in laboratory and AMT situation. While word error rates in AMT were worse than in the laboratory situation, AMT results were more sensitive to relative differences between systems. They concluded that AMT is a viable platform for synthetic speech intelligibility comparisons and boxplots was effective for identifying evaluators who performed particularly badly, while thresholding was sufficient to eliminate rogue evaluators. In addition, AMT was used to evaluate machine translation quality (Callison-Burch, 2009), speech accent (Kunath and Weinberger, 2010), and computer-generated questions (Heilman and N.A., 2010).

The evaluation comprised three sections: In section 1, speech synthesis was evaluated. Evaluators listened to synthesized speech and assigned scores for naturalness (**Naturalness**) using a five-point scale (5: completely natural – 1: completely unnatural). We asked evaluators to assign a score without considering the correctness of grammar or content. In section 2, speech-to-speech translation was evaluated. Evaluators listened to synthesized speech, then typed in the sentence; we measured their word error rate not including punctuation (**WER**). After this, evaluators assigned scores for "Adequacy" of the typed-in sentence (**S2ST-Adequacy**) using a five-point scale (5: all meaning – 1: none meaning) and assigned scores for "Fluency" of the typed-in sentence (**S2ST-Fluency**) using a five-point scale (5: flawless – 1: incomprehensible). Here, "Adequacy" indicates how much of the information from the reference translation sentence was expressed in the sentence and "Fluency" indicates that how fluent the sentence was (White et al., 1994). These definitions were provided to the evaluators. "Adequacy" and "Fluency" measures do not need bilingual evaluators; they can be evaluated by monolingual target language listeners. These measures are widely used in machine translation evaluations, e.g., conducted by NIST and IWSLT. In section 3, machine translation was evaluated. Evaluators didn't listen to synthesized speech. They read translated sentences and assigned a five-point score of "Adequacy" and "Fluency" for each sentence (**MT-Adequacy** and **MT-Fluency**).

[4] Festival http://www.festvox.org/festival/

Table 1: Example of *N*-best MT output sentences and reference sentence

N	Output sentence
Reference	We can support what you said.
1	We support what you have said.
2	We support what you said.
3	We are in favour of what you have said.
4	We support what you said about.
5	We are in favour of what you said.
6	We support what you have said about.
7	We will support what you have said.
8	We support what you have just said.
9	We support what you say.
10	We support that what you have said.

The test data comprised 100 sentences from EuroParl corpus not included in the machine translation training data. The machine translation component output the 20-best translated sentences for each input sentence, resulting in 2,000 translated sentences. The translated sentences does not include untranslated words. Table 1 shows an example of top 10-best translated sentences and reference sentence.

42 translated sentences were evaluated in each section. Evaluators were paid US$7 for the task, with the time for completion set to one hour. 150 English speakers participated in this evaluation for two days. We checked all assigned scores, and rejected some results of unreliable evaluators (e.g., the evaluators which assigned same score to almost all sentences, the evaluators which assigned scores in a very short time). As a result, we used scores assigned by 130 evaluators to analysis the impacts of machine translation and speech synthesis in the following sections.

4. Analysis between machine translation and speech synthesis

4.1. Impact of MT and WER on S2ST

First, we analyzed the impact of the translated sentences and the intelligibility of synthesized speech on S2ST. The correlation coefficients between **MT-Adequacy** and **S2ST-Adequacy** scores and between **MT-Fluency** and **S2ST-Fluency** scores were strong (0.61 and 0.68, respectively). The correlation coefficient between **WER** and **S2ST-Adequacy** score was -0.21, and the correlation coefficient between **WER** and **S2ST-Fluency** score was -0.20. These are only weak correlations. This is because **WER** averaged across all test samples was 6.49%. These results indicate that the impact of the translated sentences on S2ST is larger than the impact of the intelligibility of the synthesized speech, although the intelligibility affects the performance of S2ST.

Table 2: Correlation coefficients between **Naturalness** or **WER** and **MT** scores

	MT-Adequacy	**MT-Fluency**
Naturalness	0.12	0.24
WER	-0.17	-0.25

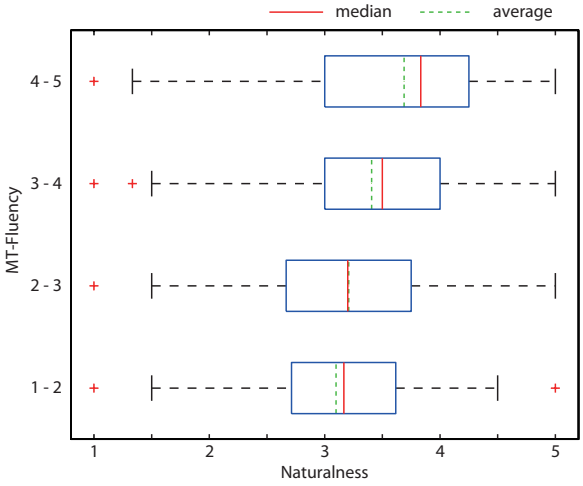

Figure 1: Boxplots of **Naturalness** score divided into four groups by **MT-Fluency** score

4.2. Impact of MT on Naturalness and WER

Next, we analyzed the impact of the translated sentences on the naturalness and intelligibility of synthesized speech. Table 2 shows the correlation coefficients between **Naturalness** and **MT** scores, and the correlation coefficients between **WER** and **MT** scores. **MT-Fluency** score correlate better with both **Naturalness** score and **WER** than **MT-Adequacy** score, while the correlation coefficients were not strong. That is, the naturalness and intelligibility of synthesized speech were more affected by the fluency of the translated sentences than by the content of them. Therefore, next we focused on the relationship between the fluency of the translation output and the synthesized speech.

Figure 1 shows boxplots of **Naturalness** score divided into four groups by **MT-Fluency** score. In this figure, the red and green lines represent the median and average scores of the groups, respectively. This figure illustrates that the median and average scores of **Naturalness** are slightly improved by increasing **MT-Fluency** score. This is presumed to be because the speech synthesis text processor (Festival, in our case) often produced incorrect full-context labels due to the errors in syntactic analysis of disfluent and ungrammatical translated sentences. In addition, the psychological effect called "Llewelyn reaction" (Yamada et al., 2005) appears to affect the results. The "Llewelyn reaction" is that evaluators perceive lower speech quality when the sentences are less fluent or the content of the sentences is less natural, even if the actual quality of synthesized speech is same. Therefore, we conclude that the speech synthesis component will tend to generate more natural speech as the translated sentences become more fluent.

Figure 2 shows the boxplots of **WER** divided into four groups by **MT-Fluency** score. From this figure, it can be seen that the median and average of **WER** improve and the variance of boxplots shrinks, with increasing **MT-Fluency** score. In particular to the most fluent group, the median score of **WER** was 0.0%. This is presumed to be because evaluators can predict the next word when the translated sentence does not include unusual words or phrases and the naturalness of synthesized speech being better when the sentences were more fluent, as previously described. Figure 3 shows the boxplots of **WER** divided into four groups by **Naturalness** score. From this figure, it can be seen that the median and average of **WER** became slightly lower when **Naturalness** score was more than three, i.e., the naturalness of synthesized speech affect the intelligibility. Therefore, the intelligibility and naturalness of synthesized speech are improved as the translated sentences become more fluent, even though all sentences are synthesized by the same system.

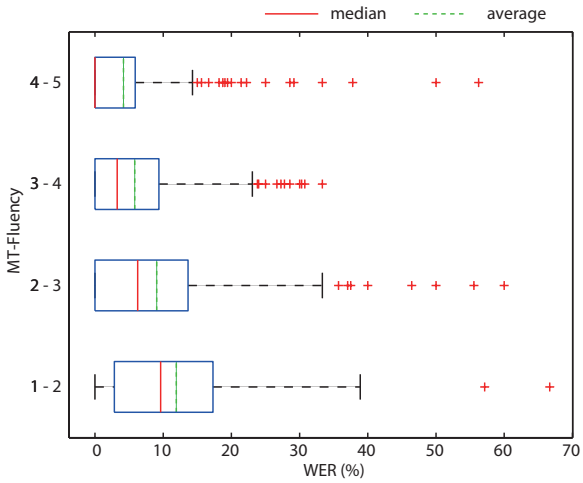

Figure 2: Boxplots of **WER** score divided into four groups by **MT-Fluency** score

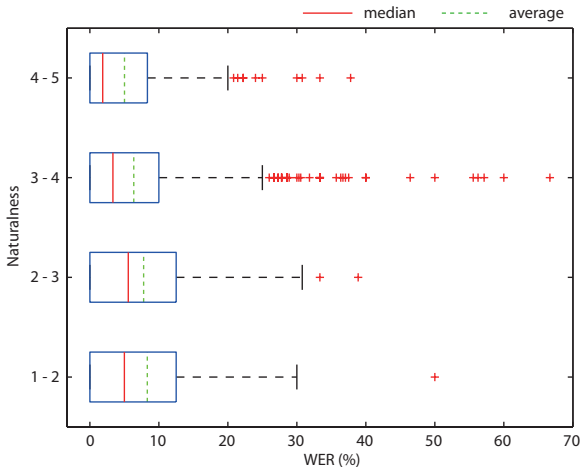

Figure 3: Boxplots of **WER** score divided into four groups by **Naturalness** score

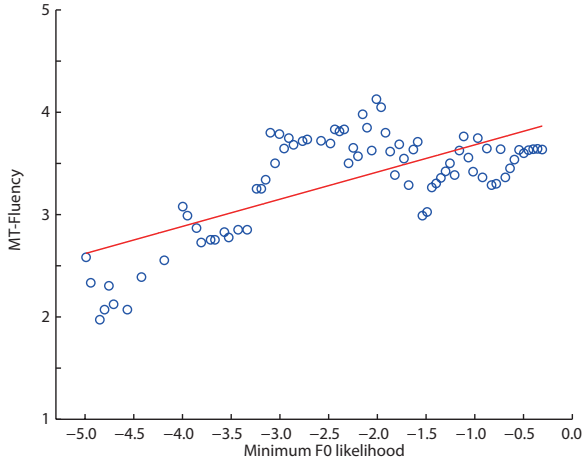

Figure 4: Correlation between bin-averaged **MT-Fluency** scores and minimum F_0 likelihood ($r = 0.70$, $p < 0.01$)

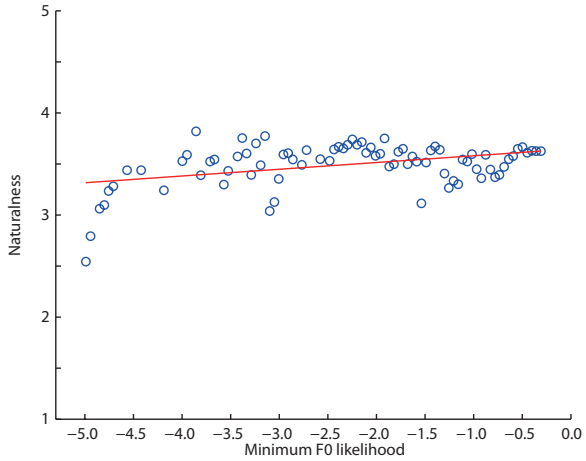

Figure 5: Correlation between bin-averaged **Naturalness** scores and minimum F_0 likelihood ($r = 0.40$, $p < 0.01$)

5. Prediction of MT Fluency and Naturalness

5.1. Correlation with score of speech synthesizer

We have shown that the naturalness and intelligibility of the synthesized speech are strongly affected by the fluency of sentences. Therefore, we looked for the objective measures which can predict the fluency of translated sentences. Synthesized speech data likelihood (i.e., likelihood of generated speech parameters) is a measure of synthesized speech quality in HMM-based speech synthesis. The likelihood represents the fit of the model to the data. When the synthesized speech data likelihood is small, the error of speech synthesizer is increased and the synthesized speech would become low quality. Therefore, we computed the correlation coefficients with synthesized speech data likelihood. Since various sentences were evaluated in this evaluation and the lengths of synthesized speech were different, it was difficult to compare synthesized speech data likelihood directly. Therefore, we used a minimum frame likelihood which would represent the lowest local quality of synthesized speech for computing correlation. Then, we found that the minimum frame likelihood of F_0 correlates well with the subjective scores **MT-Fluency** and **Naturalness**.

Figure 4 shows the bin-averaged **MT-Fluency** score and minimum frame likelihood of F_0, and Figure 5 shows the bin-averaged **Naturalness** score and minimum frame likelihood of F_0. The correlation coefficients were 0.70 ($p < 0.01$) and 0.40 ($p < 0.01$), respectively. We will elaborate how to compute bin-averaged score and its motivations behind in Section 5.3. Although the minimum frame likelihood of all features did not correlate with **MT-Fluency** and **Naturalness** scores ($r = -0.14$, $p = 0.23$ and $r = -0.18$, $p = 0.23$, respectively), a strong correlation was observed by focusing on F_0. In particular, the minimum frame likelihood of F_0 correlated well with **MT-Fluency** score. Thus, it is indicated that the fluency of translated sentences affects prosody of the speech synthesizer. This agrees with the result that **Naturalness** score is improved by increasing **MT-Fluency** score, as shown in Figure 1. These results imply that the minimum frame likelihood of F_0 is one of appropriate features for measuring the average fluency of translated sentences and the average naturalness of synthesized speech.

5.2. Correlation with score of text processor

We presumed that the text processor included in speech synthesis often produced incorrect full-context labels due to the errors in syntactic analysis of disfluent and ungrammatical translated sentences. Therefore, we computed the correlation coefficients between **MT-Fluency** and text processor scores. Here, we used an averaged probability of POS tagging per word as a text processor score.

Figure 6 shows the bin-averaged **MT-Fluency** and text processor scores, and Figure 7 shows the bin-averaged **Naturalness** and text processor scores. When the score of text processor was small, the bin-averaged **MT-Fluency** score was widely distributed. This is because that the number of samples being small score of text processor was insufficient. The correlation coefficient between the bin-averaged **MT-Fluency** and text processor scores was 0.43 ($p < 0.01$). The fluency of translated sentences weakly affects syntactic analysis and POS tagging. The correlation coefficient between the bin-averaged **Naturalness** score and score of text processor was 0.28 ($p = 0.01$), and there was not a strong correlation. This is because the score of text processor represents the complexity of POS tagging rather than the number of errors in syntactic analysis. These results suggest that the score of text processor may be optionally used for measuring the average perceived fluency of translated sentences, although it is difficult to predict the naturalness of synthesized speech.

5.3. Correlation between MT Fluency and word N-gram scores

It is well known in the field of machine translation that the fluency of translated sentences can be improved by using long-span word-level N-grams. Therefore, we computed the correlation coefficient between **MT-Fluency** and word N-gram scores. Here, we used an average probability per word (perplexity) as a word N-gram score. Perplexity is a measure of average branching factor and can be used to measure how well an N-gram

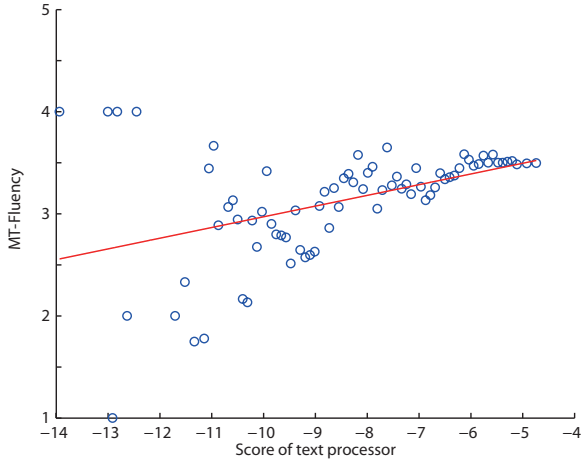

Figure 6: Correlation between bin-averaged **MT-Fluency** and text processor score ($r = 0.43$, $p < 0.01$)

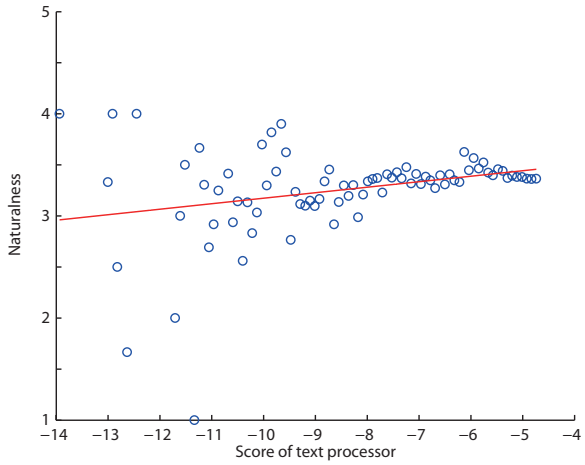

Figure 7: Correlation between bin-averaged **Naturalness** and text processor score ($r = 0.28$, $p = 0.01$)

model predicts the next word. The word N-gram models we used were created using the SRILM toolkit (Stolcke, 2002), from the same English sentences used for training the machine translation component. Kneser-Ney smoothing (Kneser and Ney, 1995) was employed.

Table 3 shows the correlation coefficients between **MT-Fluency** and word N-gram score. It can be seen from Table 3 that the word N-gram score directly correlate well with **MT-Fluency** score even on raw data, as N is larger and that the word 5-gram gave the strongest correlation coefficient of 0.44 ($p < 0.01$). Figure 8 illustrates the scatter plot of **MT-Fluency** and word 5-gram score. Although the word N-gram scores correlate with **MT-Fluency** on the raw data, what we want to have stronger interests in more is averaged rough tendency across sentences and evauators. Therefore, **MT-Fluency** scores were divided into 100 bins according to the word 5-gram score and subsequently average **MT-Fluency** scores for each bin were computed. In Figure 9, the bin-averaged **MT-Fluency** and word 5-gram scores are shown, and the regression line is illustrated by the red line. Then, the correlation coefficient was

Table 3: Table of correlation coefficients between **MT-Fluency** and word N-gram scores

1-gram	2-gram	3-gram	4-gram	5-gram
0.28	0.39	0.42	0.43	0.44

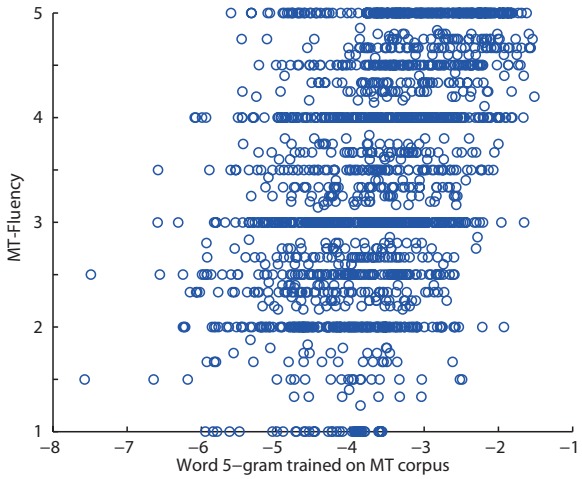

Figure 8: Scatter plot of **MT-Fluency** and word 5-gram scores

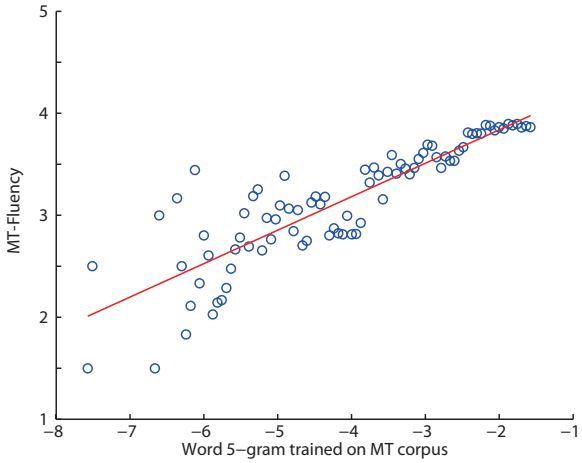

Figure 9: Correlation between bin-averaged **MT-Fluency** and word 5-gram scores ($r = 0.87$, $p < 0.01$)

0.87 ($p < 0.01$). Since the **MT-Fluency** on the raw data score varies depending on the translated sentences and the evaluators, it was confirmed that the strong correlation was shown by averaging the **MT-Fluency** scores. This result indicates that the word 5-gram score is the most appropriate feature for measuring the average perceived fluency of translated sentences in our experiments.

5.4. Correlation between Naturalness and phoneme N-gram scores

P.563 is an objective measure for predicting the quality of natural speech in telecommunication applications (Malfait et al., 2006). However, we found no correlation between **Naturalness** score and P.563 ($r = 0.03$, $p = 0.24$ on raw data).

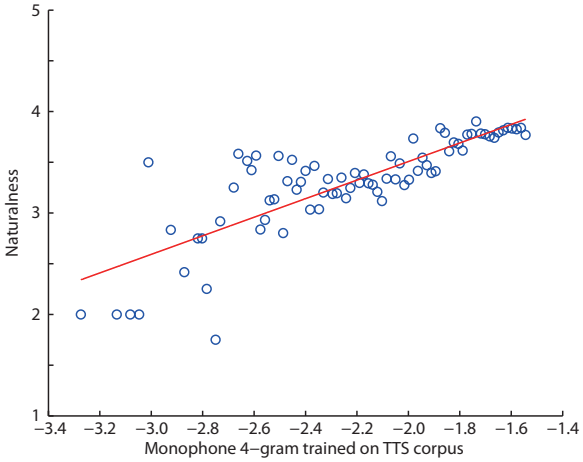

Figure 10: Correlation between bin-averaged **Naturalness** scores and monophone 4-gram scores ($r = 0.81$, $p < 0.01$)

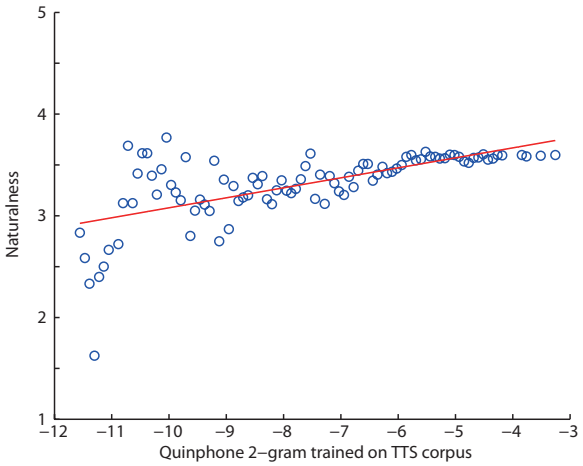

Figure 11: Correlation between bin-averaged **Naturalness** scores and quinphone 2-gram scores ($r = 0.64$, $p < 0.01$)

So, we looked for correlations with other objective measures. It is well known that HMM-based speech synthesis systems generally produce better quality speech when the input sentence is in-domain (i.e., similar to sentences found in the training data). Therefore, we computed the correlation coefficient between **Naturalness** and N-gram scores of the sentence being synthesized; the N-gram score is a measure of the coverage provided by the training data for that particular sentence. Since the corpus used for training the speech synthesis component was significantly smaller than one used for training the machine translation component, the use of word N-gram estimated from the speech synthesis corpus was difficult. Therefore, we used the phoneme N-gram model estimated from the speech synthesis corpus. This would represent segmental quality of synthetic speech to some extend.

Figure 10 shows the bin-averaged **Naturalness** and monophone 4-gram scores, and Figure 11 shows the bin-averaged **Naturalness** and quinphone 2-gram scores. The correlation coefficients were 0.81 ($p < 0.01$) and 0.64 ($p < 0.01$), respectively. The correlation between bin-averaged **Naturalness** and phoneme N-gram scores was strong. These results suggest that the monophone 4-gram and/or quinphone 2-gram scores are good measures for predicting a rough trend of naturalness of synthesized speech.

The ability to predict average naturalness of synthesized speech before generating the speech could be used in other applications, such as sentence selection (as in this work, or in natural language generation with speech output), voice selection before generating speech. We hope to investigate this further in the future.

5.5. Summary of analyses

The naturalness and intelligibility of synthesized speech in the S2ST system are improved as the translated sentences become more fluent, even if all sentences are synthesized by the same system. We found that perceived fluency of the translated texts correrates well the minimum F_0 likelihood. This means that prosody of synthetic speech is affected by the fluency of the translated texts. We also found that long-span word N-gram and phoneme N-gram scores correlate well with the fluency of translated sentence and the naturalness of synthesized speech, respectively.

6. Conclusion

This paper has provided an analysis of the impacts of machine translation and speech synthesis on speech-to-speech translation. We have shown that the fluency of the translated sentences strongly affected the quality of synthesized speech. The naturalness and intelligibility of synthesized speech are improved as the translated sentence become more fluent. Therefore, the fluency is one of the most important factor for speech synthesis systems in the S2ST systems. We found that perceived fluency of the translated texts correrates well the minimum F_0 likelihood, meaning that prosody of synthetic speech is affected by the fluency of the translated texts. In addition, we have looked for the objective measures which can predict the fluency of translated sentences and the naturalness of synthesized speech. Results of analyses showed that the long-span word N-gram and phoneme N-gram scores correlate well with the fluency of translated sentence and the naturalness of synthesized speech, respectively. These objective measures can be used for predicting a rough trend of the fluency and naturalness before the S2ST system synthesise the translated sentences. However, we have not found features which correlate with "Adequacy" well. Our future work will include more detail analyses of the impact of machine translation and speech synthesis and investigations into the integration methods of machine translation and speech synthesis using features which correlate with subjective scores.

7. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME

project http://www.emime.org) and the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communication, Japan. A part of this research was supported by JSPS (Japan Society for the Promotion of Science) Research Fellowships for Young Scientists.

References

Boidin, C., Rieser, V., Plas, L., Lemon, O., Chevelu, J., 2009. Predicting how it sounds: Re-ranking dialogue prompts based on TTS quality for adaptive spoken dialogue systems. Proceedings of Interspeech 2009, 2487–2490.

Bulyko, I.and Ostendorf, M., 2002. Efficient integrated response generation from multiple target using weighted finite state transducers. Computer Speech and Language 16, 533–550.

Callison-Burch, C., 2009. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. Proceedings of EMNLP, 286–295.

Callison-Burch, C., Dredze, M., 2010. Creating speech and language data with amazon's mechanical turk. Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 1–12.

Gispert, A., Virpioja, S., Kurimo, M., Byrne, W., 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. Proceedings of NAACL-HLT 2009, 73–76.

Heilman, M., N.A., S., 2010. Rating computer-generated questions with mechanical turk. Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 35–40.

Iglesias, G., Gispert, A., Banga, E., Byrne, W., 2009. Hierarchical phrase-based translation with weighted finite state transducers. Proceedings of NAACL-HLT 2009, 433–441.

Kawahara, H., Masuda-Katsuse, I., Cheveigne, A., 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication 27, 187–207.

Kneser, R., Ney, H., 1995. Improved backing-off for m-gram language model. Proceedings of ICASSP 1995, 181–184.

Koehn, P., 2005. Europarl: A parallel corpus for statistical machine translation. Proceedings of MT Summit, 79–86.

Kunath, S., Weinberger, S., 2010. The wisdom of the crowd's ear: speech accent rating and annotation with amazon mechanical turk. Proceedings of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 168–171.

Malfait, L., Berger, J., Kastner, M., 2006. P.563 – The ITU-T standard for signal-ended speech quality assesment. IEEE Transactions on Audio, Speech and Language Processing 14 (6), 1924–1934.

Nakatsu, C., White, M., 2006. Learning to say it well: Reranking realizations by predicted synthesis quality. Proceedings of ACL 2006, 1113–1120.

Ney, H., 1999. Speech translation: coupling of recognition and translation. Proceedings of ICASSP 1999, 1149–1152.

Noth, E., Batliner, A., Kießling, A., R., K., Niemann, H., 2000. VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system. IEEE Transactions on Speech and Audio Processing 8 (5), 519–532.

Parlikar, A., Black, A., Vogel, S., 2010. Improving speech synthesis of machine translation output. Proceedings of Interspeech 2010, 194–197.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A., 2008. Cheap and fast–but is it good?: evaluating non-expert annotations for natural language tasks. Proceedings of EMNLP, 254–263.

Stolcke, A., 2002. SRILM – An extensible language model toolkit. Proceedings of ICSLP 2002, 901–904.

Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. Proceedings of ICASSP 1999, 229–232.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. Proceedings of ICASSP 2000, 936–939.

Vidal, E., 1997. Finite-State Speech-to-Speech Translation. Proceedings of ICASSP 1997, 111–114.

White, J., O'Connell, T., O'Mara, F., 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proceedings of AMTA, 193–205.

Wolters, M., Isaac, K., Renals, S., 2010. Evaluating speech synthesis intelligibility using amazon mechanical turk. Proceedings of SSW7, 136–141.

Wu, Y., Nankaku, Y., Tokuda, K., 2009. State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. Proceedings of Interspeech 2009, 528–531.

Yamada, S., Kodama, S., Matsuoka, T., Araki, H., Murakami, Y., Takano, O., Sakamoto, Y., 2005. A report on the machine translation market in Japan. Proceedings of MT Summit X, 55–62.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proceedings of Eurospeech 1999, 2347–2350.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 2004. Hidden semi-Markov model based speech synthesis. Proceedings of ICSLP, 1185–1180.